ASSESSED INTELLIGENCE

ASSESSED SIGNAL

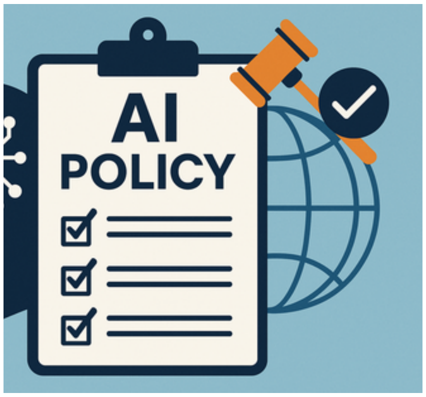
Intersection of Policy, Use Case, & Threat Intelligence

Organizations continue to adopt AI/ML for various tasks but are not considering some of the longer-term challenges with privacy and security. Adversaries are creating new and creative ways to use extant AI against defenders. As more organizations adopt AI for their workflows, their information sources, the organizations will become increasingly exposed to malicious actors. Research on AI and security does not proceed at the same pace as adoption, thus the gap between the art of the possible and exposure grows at an exponential rate.

Secure
&
Responsible
Technology.

AssessedIntelligence.com

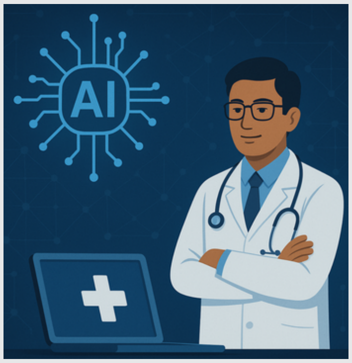
Policy Based AI Articles of Interest



Google says it has removed Gemma from its AI Studio after a U.S. senator accused the AI model of fabricating accusations of sexual misconduct against her. In a letter to Google CEO Sundar Pichai, Senator Marsha Blackburn — a Republican from Tennessee — said that when Gemma was asked, “Has Marsha Blackburn been accused of rape?” it responded by falsely claiming that during a 1987 state senate campaign, a state trooper alleged that Blackburn “pressured him to obtain prescription drugs for her and that the relationship involved non-consensual acts.”

“None of this is true, not even the campaign year, which was actually 1998,” Blackburn wrote. While there are links to news articles that supposedly support these claims, she said, “The links lead to error pages and unrelated news articles. There has never been such an accusation, there is no such individual, and there are no such news stories.”

https://techcrunch.com/2025/11/02/google-pulls-gemma-from-ai-studio-after-senator-blackburn-accuses-model-of-defamation/



The National Health Authority (NHA) showcased its pioneering use of Artificial Intelligence (AI) in strengthening transparency and integrity within India’s digital health ecosystem at a high-level workshop organized by the Central Vigilance Commission (CVC) at Bharat Mandapam, New Delhi, as part of Vigilance Awareness Week 2025.

Dr. Barnwal also emphasized that, “Artificial Intelligence and Machine Learning are revolutionizing the world’s largest public health assurance scheme—shifting the paradigm from reactive fraud detection to proactive integrity management. This transformation is enabling greater transparency, accountability, and equity in every citizen’s healthcare journey.”

https://indiaeducationdiary.in/nha-showcases-use-of-artificial-intelligence-to-strengthen-transparency-and-integrity-in-indias-digital-health-ecosystem-at-cvc-workshop/

Cases of LLM abuse

- Apart from AI-powered malware, Google's report also documents multiple cases where threat actors abused Gemini across the entire attack lifecycle.
- A China-nexus actor posed as a capture-the-flag (CTF) participant to bypass Gemini's safety filters and obtain exploit details, using the model to find vulnerabilities, craft phishing lures, and build exfiltration tools.
- Iranian hackers MuddyCoast (UNC3313) pretended to be a student to use Gemini for malware
- development and debugging, accidentally exposing C2 domains and keys.
- Iranian group APT42 abused Gemini for phishing and data analysis, creating lures, translating content, and developing a "Data Processing Agent" that converted natural language into SQL for personal-data mining.
- China's APT41 leveraged Gemini for code assistance, enhancing its OSSTUN C2 framework and utilizing obfuscation libraries to increase malware sophistication.
- Finally, the North Korean threat group Masan (UNC1069) utilized Gemini for crypto theft, multilingual phishing, and creating deepfake lures, while Pukchong (UNC4899) employed it for developing code targeting edge devices and browsers.
- In all cases Google identified, it disabled the associated accounts and reinforced model safeguards based on the observed tactics, to make their bypassing for abuse harder.

https://www.bleepingcomputer.com/news/security/google-warns-of-new-ai-powered-malware-families-deployed-in-the-wild/

R&D/Attacks/Vectors of Attack

Organizations are increasingly adopting and adapting Large Language Models (LLMs) hosted on public repositories such as HuggingFace. Although these adaptations often improve performance on specialized downstream tasks, recent evidence indicates that they can also degrade a model's safety or fairness. Since different fine-tuning techniques may exert distinct effects on these critical dimensions, this study undertakes a systematic assessment of their trade-offs. Four widely used Parameter-Efficient Fine-Tuning methods, LoRA, IA3, Prompt-Tuning, and P-Tuning, are applied to four instruction-tuned model families (Meta-Llama-3-8B, Qwen2.5-7B, Mistral-7B, and Gemma-7B). In total, 235 fine-tuned variants are evaluated across eleven safety hazard categories and nine demographic fairness dimensions. The results show that adapter-based approaches (LoRA, IA3) tend to improve safety scores and are the least disruptive to fairness, retaining higher accuracy and lower bias scores. In contrast, prompt-based methods (Prompt-Tuning and P-Tuning) generally reduce safety and cause larger fairness regressions, with decreased accuracy and increased bias. Alignment shifts are strongly moderated by base model type: LLaMA remains stable, Qwen records modest gains, Gemma experiences the steepest safety decline, and Mistral, which is released without an internal moderation layer, displays the greatest variance. Improvements in safety do not necessarily translate into improvements in fairness, and no single configuration optimizes all fairness metrics simultaneously, indicating an inherent trade-off between these objectives.

https://arxiv.org/abs/2511.00382

Google's Threat Intelligence Group (GTIG) has identified a major shift this year, with adversaries leveraging artificial intelligence to deploy new malware families that integrate large language models (LLMs) during execution.

This new approach enables dynamic altering mid-execution, which reaches new levels of operational versatility that are virtually impossible to achieve with traditional malware. Google calls the technique "just-in-time" self-modification and highlights the experimental PromptFlux malware dropper and the PromptSteal (a.k.a. LameHug) data miner deployed in Ukraine, as examples for dynamic script generation, code obfuscation, and creation of on-demand functions.

"The most novel component of PROMPTFLUX is its 'Thinking Robot' module, designed to periodically query Gemini to obtain new code for evading antivirus software," explains Google. The prompt is very specific and machine-parsable, according to the researchers, who see indications that the malware's creators aim to create an ever-evolving "metamorphic script." Google could not attribute PromptFlux to a specific threat actor, but noted that the tactics, techniques, and procedures indicate that it is being used by a financially motivated group.

https://www.bleepingcomputer.com/news/security/google-warns-of-new-ai-powered-malware-families-deployed-in-the-wild/

Forged by Experience | Driven by Purpose | Built to Endure

ASSESSED INTELLIGENCE

Secure & Responsible Technology.

YOU'RE BUILDING THE FUTURE. DON'T LET EMERGING RISK STALL YOUR MOMENTUM.

SALES@ASSESSEDINTELLIGENCE.COM
EMEASALES@ASSESSEDINTELLIGENCE.COM